THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 56, No. 9, November 1977
Printed in U.S.A.

Criteria for Determining if a High-Order Digital Filter Using Saturation Arithmetic Is Free of Overflow Oscillations

By DEBASIS MITRA

(Manuscript received May 2, 1977)

Recently we found that, among recursive digital filters using saturation arithmetic to contend with overflow, a fundamental difference exists between second and higher order filters: the latter may sustain large-amplitude overflow oscillations. In this paper we have derived a new criterion expressly designed for determining when a given high-order recursive system using saturation arithmetic is free of overflow oscillations. The new criterion, which is easy to use, follows from this result: we associate with the given system two trigonometric polynomials in θ of degree equal to the order of the given system; if any linear combination of the polynomials with nonnegative weights is positive for all θ in $[0,\pi]$ then the system is free of all nontrivial periodic oscillations. We prove that the new criterion subsumes certain well-known criteria, such as Tsypkin's criterion, from the literature on nonlinear systems. To illustrate, three classes of special systems are investigated, and in each case the new criterion gives substantial improvements. Finally, the new test is applied in the synthesis of high order sections for a realistic eighth-order system.

I. INTRODUCTION

Recently[1] we made the unexpected observation that, among recursive filters employing saturation arithmetic, a fundamental difference exists between second and higher order filters, namely, the latter may sustain large-amplitude overflow oscillations. This observation proved to be timely since it coincided with the awareness that economies of scale coupled with various recent developments make highly attractive the use of high-order sections in filter realizations. It has also come to light that the problem of possible large-scale oscillations is of interest not only in data filtering but in other areas where the natural structure is a

high-order recursive system, e.g., code converters (DPCM → PCM) and speech synthesizers.

The economies of scale derive from the fact that the overflow detection and correction circuits, an expensive part of present-day filters, are as many as the number of sections employed; thus if a realization is composed of fourth-order sections rather than the conventional second-order sections, then the number of such circuits may be expected to be halved. The recent developments alluded to earlier refer to the almost simultaneous developments of inexpensive, lower-power-consuming semiconductor read-only memories, and the concept of distributed arithmetic blocks[2,3] in which ROMs are used to implement digital filters. In a pioneering study R. B. Kieburtz[4] recently estimated that in a particular application a saving of about 30 percent in parts may be achieved over the conventional design through the use of fourth-order sections using saturation arithmetic and implemented by ROMs.

Thus there is much to be gained if high-order sections can be used, and for this to happen it is first necessary to ensure that the highly destructive overflow oscillations are not present in a particular design. It is apparent that there is a useful role for an effective criterion for delineating stable systems employing saturation arithmetic. It is possible to conceive of the situation where such a criterion is incorporated in the early design, i.e., the criterion is introduced as a constraint in the approximation problem. The other possibility of an arithmetic different from saturation arithmetic to contend with overflow is not pursued in this paper.

There do exist many such criteria in the literature on the stability of a class of nonlinear feedback systems (i.e., the systems in the Lurie problem[24]) of which the one under consideration here is a member[5–11]; the reader may consult Ref. 8 for a comparative evaluation of some of these criteria. These criteria are in some sense generalizations of Nyquist's criterion for linear feedback systems. The reader will find in Sec. 5.2 a statement, in the context of the present problem, of Tsypkin's criterion and the discrete circle criterion, two well-known examples of such criteria. Unfortunately it is known that when the nonlinearity in the system is the one associated with saturation arithmetic then these criteria, including Barkin's criterion,[7,8] are of limited utility since they are excessively pessimistic. Examples to this effect may be found here. Also telling is the fact that these criteria do not predict that all second-order systems using saturation arithmetic are free of oscillations,[8] a fact proven in Refs. 12–14 by arguments special to second-order systems. This is not totally unexpected in view of the fact that the systems-theoretical criteria apply to large classes of systems and nonlinearities and, concomitantly, use relatively little information (restricted to the sector information, symmetry, and monotonicity) about the nonlinearity.

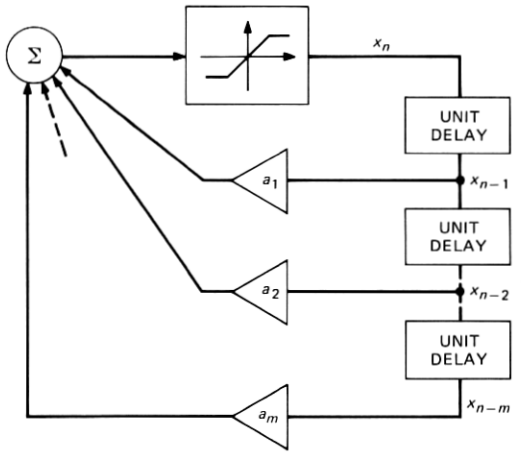

Fig. 1—Schematic of unforced high-order filter employing saturation arithmetic.

In this paper we have derived new criteria expressly designed for the system employing saturation arithmetic. Underlying the new criteria is the observation that certain unique passivity conditions are operative in the case of saturation arithmetic. Both the observations regarding the passivity conditions as well as the technique we use for deriving the criterion are believed to be new. The main ingredient in the derivation is the observation that the expressions associated with the passivity conditions in periodic solutions possess remarkable structure; namely, they are quadratic forms involving circulant matrices.

The systems considered in this paper are of the form (see Figs. 1 and 2)

$$x_n = F\left(\sum_{j=1}^{m} a_j x_{n-j}\right), \quad n = 0,1,\ldots \tag{1}$$

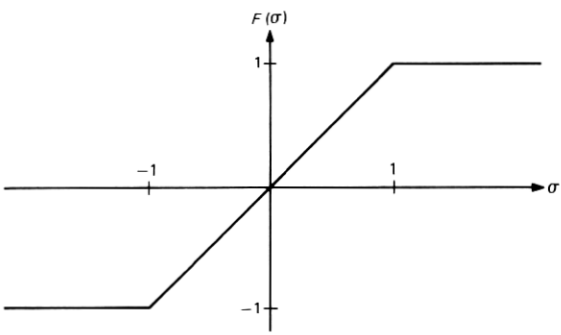

Fig. 2—The saturation arithmetic nonlinearity.

where m is the order of the system, $\{a_j\}$ are the coefficients, and $F(\cdot)$ is a nonlinear function associated with saturation arithmetic, namely

$$F(\sigma) = \sigma \quad \text{if } |\sigma| \leq 1$$
$$= \operatorname{sgn} \sigma \text{ if } |\sigma| \geq 1 \tag{2}$$

It is tacitly assumed that the underlying linear system in eq. (1) is absolutely stable, i.e.,

$$\lambda^m - \sum_{j=1}^{m} a_j \lambda^{m-j} \neq 0 \text{ for } all \ |\lambda| \geq 1. \tag{3}$$

Thus any nontrivial solution of eq. (1) will necessarily have either 1 or -1 as an element and consequently such solutions are referred to as overflow oscillations. Note that we are following convention in ignoring quantization effects in the description of the filter in eq. (1); in investigations of large-scale oscillations it is natural to focus on the gross nonlinearity.

The main result of this paper (Sec. 2.4) is simply stated: for a given system of order m with coefficients $\{a_j\}$, we associate two polynomials of degree m in $\cos \theta$, namely,

$$p_1(\theta) = 1 - \cos \theta - \sum_{j=1}^{m} a_j \{\cos j\theta - \cos (j-1)\theta\} \tag{4}$$

and

$$p_2(\theta) = 1 + \cos \theta - \sum_{j=1}^{m} a_j \{\cos j\theta + \cos (j-1)\theta\}. \tag{5}$$

If any linear combination of the polynomials with nonnegative weights is positive for all θ in $[0,\pi]$ then the system in eq. (1) with arbitrary initial conditions does not admit any nontrivial periodic solutions. Certain generalizations of this criterion are derived in Sec. VI.

In Sec. 2.5 ("How To Use The New Test") we show that the criterion may be used in a straightforward manner by one of two methods. The first method calls for plots of $p_1(\theta)$, $p_2(\theta)$ and $p_1(\theta)/p_2(\theta)$ for θ in $[0,\pi]$. The second requires the consistency of a set of linear inequalities to be checked. The second method may also be used for the generalized criterion in Sec. VI.

In Sec. III we examine three classes of special systems in detail. The results for the following canonical example[1] are typical: in a fourth-order system the poles are taken to be all real and repeated at ρ where $|\rho| < 1$. For $|\rho| \geq 0.669$ overflow oscillations are proven to exist. Tsypkin's criterion and the circle criterion guarantee the absence of oscillations for $|\rho| \leq 0.384$. The new test guarantees the absence of oscillations for $|\rho| \leq 0.610$.

In Sec. IV we apply the test to an eighth-order filter in the TDM-FDM translator, an extensively studied application of digital filtering. We find that this system can be lumped into two fourth-order sections employing saturation arithmetic, neither of which can sustain overflow oscillations. Both sections fail Tsypkin's criterion and the circle criterion tests.

Finally, in Sec. V we prove that the new criterion (*i*) easily gives the well-known result that overflow oscillations do not exist in second-order sections, and (*ii*) subsumes Tsypkin's criterion.

II. THE CRITERION

2.1 Passivity properties

On account of the special form of the nonlinearity F the system in eq. (1) possesses certain simply stated but important properties which we interpret as passivity properties. The criterion we derive is a direct consequence of these properties.

We may write eq. (1) as a linear system with a forcing sequence present by defining

$$e_n \triangleq F\left(\sum_1^m a_j x_{n-j}\right) - \left(\sum_1^m a_j x_{n-j}\right), \quad n = 0,1,\ldots \tag{6}$$

so that, from eq. (1),

$$x_n = \sum_{j=1}^m a_j x_{n-j} + e_n, \quad n = 0,1,\ldots \tag{7}$$

The above nonhomogeneous linear recursion is used throughout the paper. Our procedure is to translate the important features of $F(\cdot)$ into tractable constraints on $\{e_n\}$.

Every solution of eq. (1) possesses the following properties.

Proposition 1:

$$e_n(x_n - x_{n-1}) \leqslant 0, \quad n = 1,2,\ldots \tag{8}$$

and

$$e_n(x_n + x_{n-1}) \leqslant 0, \quad n = 1,2,\ldots \tag{9}$$

For proof we observe that if $|\Sigma\, a_j x_{n-j}| \leqslant 1$ then $e_n = 0$ and both conditions are obviously valid. If

$$\sum_{j=1}^m a_j x_{n-j} > 1 \quad \text{then} \quad x_n = 1 \text{ from (1)}$$

$$\text{and} \quad e_n < 0 \text{ from (6)}$$

and, of course, $|x_{n-1}| \leqslant 1$. Thus, in this case eqs. (8) and (9) are valid.

Similarly, if $\Sigma a_j x_{n-j} < -1$ then $x_n = -1$, $e_n > 0$ and as $|x_{n-1}| \leqslant 1$, the relations remain valid.

Equation (8) may be viewed as stating that the forcing term e_n has opposite sign from $(x_n - x_{n-1})$ which we may interpret as the discrete analog of "velocity." Thus locally the forcing term acts to reduce the velocity. Similarly interpreting $(x_n + x_{n-1})/2$ as the local "distance," eq. (9) states that locally the forcing term also acts to reduce the distance. For these reasons we view eqs. (8) and (9) as passivity properties.

Note that the two above conditions imply the following weaker condition:

$$e_n x_n \leqslant 0, \quad n = 1,2,\ldots \tag{10}$$

We find upon reflection that the latter condition is completely equivalent to the nonlinearity F in eq. (1) lying in the sector bounded by lines of slopes 0 and 1:

$$0 \leqslant F(\sigma)/\sigma \leqslant 1 \quad \text{for } all\ \sigma. \tag{11}$$

This sector information is exploited in various criteria[5–11] but the additional information in eqs. (8) and (9) is not.

2.2 Equations for an oscillatory solution

We state the equations associated with every oscillatory solution of period N of the nonhomogeneous recursion in eq. (7). We find that the equations in matrix form involve a circulant matrix. We put into perspective some well-known results on circulants which are assembled in Appendix A.

A periodic solution of eq (7) with period N has associated with it the following set of N equations involving the coefficients $\{a_j\}$, the elements of the solution $X_1, X_2, \ldots, X_N$ and the corresponding forcing terms $E_1, E_2, \ldots, E_N$:

$$\begin{aligned} X_1 &= a_1 X_N + a_2 X_{N-1} + \cdots + a_m X_{N-m+1} + E_1 \\ X_2 &= a_1 X_1 + a_2 X_N + \cdots + a_m X_{N-m+2} + E_2 \\ &\vdots \\ X_N &= a_1 X_{N-1} + a_2 X_{N-2} + \cdots + a_m X_{N-m} + E_N \end{aligned} \tag{12}$$

In matrix form,

$$\mathbf{MX} = \mathbf{E}. \tag{13}$$

The interesting feature of the $N \times N$ matrix $\mathbf{M}$ is that it is a circulant[15–17] since it is the following polynomial in the primitive $N \times N$ circulant matrix $\mathbf{P}$ (see Appendix A for definition of $\mathbf{P}$):

$$\mathbf{M} = \mathbf{I} - \sum_{j=1}^{m} a_j \mathbf{P}^j \tag{14}$$

The circulant matrices have been extensively studied in the past and we are in the fortunate position of knowing a great deal of their eigenstructure.† In particular, the eigenvalues of **M** are

$$1 - \sum_{j=1}^{m} a_j e^{-ijk2\pi/N}, \quad k = 1,2,\ldots,N \tag{15}$$

The eigenvectors of circulants are also known. The following remarkable property of the eigenvectors of circulants is of utmost importance in the paper (see next section): all $N \times N$ circulants have an identical set of eigenvectors, i.e., the eigenvectors do not depend on the constituents of the matrix. Thus, the eigenvectors of any $N \times N$ circulant are‡:

$$\mathbf{u}_k = \frac{1}{\sqrt{N}} \{e^{-ik(2\pi/N)}, e^{-i2k(2\pi/N)}, \ldots, e^{-i(N-1)k(2\pi/N)}, 1\}^* \tag{16}$$

$k = 1,2,\ldots,N$. Thus the real and imaginary components of the elements of each eigenvector are sequences of equiseparated samples of a sine function. Although these are complex vectors and circulants are not generally symmetric, the eigenvectors of circulants share an important property with eigenvectors of symmetric matrices in that they form an orthonormal set, i.e.,

$$\mathbf{u}_k^* \mathbf{u}_l = \delta_{kl}. \tag{17}$$

In matrix notation,

$$\mathbf{U}^*\mathbf{U} = \mathbf{I} \tag{18}$$

where the eigenvectors $\{\mathbf{u}_k\}$ have been arranged as columns of the matrix **U**.

2.3 Another representation of the passivity properties

We combine the above information with the passivity properties stated in Proposition 1 to obtain a compact and useful representation of the passivity properties that are valid if an oscillatory solution to (1) exists. As in the preceding section an oscillatory solution is assumed to be of period N.

Note that we may write

$$(X_N, X_1, \ldots, X_{N-1}) = (X_1, X_2, \ldots, X_N)\mathbf{P}' = \mathbf{X}'\mathbf{P}' \tag{19}$$

† Recently we have had another occasion[18] to use the eigenstructure of the matrix M. Willson[19] investigates the matrix M from a different angle.

‡ We denote the conjugate transpose by the superscript *. In the case of real matrices it is also denoted by the superscript '.

where $\mathbf{P}$ is the primitive $N \times N$ circulant. Thus

$$\sum_{n=1}^{N} E_n(X_n - X_{n-1}) = \mathbf{X}'(\mathbf{I} - \mathbf{P}')\mathbf{E}$$

$$= \mathbf{X}'(\mathbf{I} - \mathbf{P}')\mathbf{M}\mathbf{X}, \quad \text{from (13)}$$

$$= \mathbf{X}'\left[\mathbf{I} - \sum_{j=1}^{m} a_j \mathbf{P}^j - \mathbf{P}^{N-1} + \sum_{j=1}^{m} a_j \mathbf{P}^{j-1}\right]\mathbf{X} \tag{20}$$

where in the final step we have used, in addition to the expression for $\mathbf{M}$, the relations $\mathbf{P}' = \mathbf{P}^{N-1}$ and $\mathbf{P}^N = \mathbf{I}$. The key observation about eq. (20) is that the matrix there, being a polynomial in $\mathbf{P}$, is a circulant.

We undertake a convenient change of coordinates to diagonalize the matrix in eq. (20). Let

$$\mathbf{Z} \triangleq \mathbf{U}^*\mathbf{X} \tag{21}$$

where $\mathbf{U}$ is, as in Sec. 2.2, the unitary matrix of eigenvectors of $N \times N$ circulants. Denoting the known eigenvalues (see Appendix A) of the matrix in eq. (20) by μ_k, k = 1,. . . ,N, we obtain

$$\sum_{n=1}^{N} E_n(X_n - X_{n-1}) = \sum_{k=1}^{N} |Z_k|^2 \operatorname{Re} \mu_k \tag{22}$$

Now,

$$\operatorname{Re} \mu_k = 1 - \cos\{k(2\pi/N)\} - \sum_{j=1}^{m} a_j[\cos\{jk2\pi/N\} - \cos\{(j-1)k2\pi/N\}] \tag{23}$$

To put eqs. (22) and (23) into the most convenient form, define the polynomial $p_1(\cdot)$ where

$$p_1(\theta) \triangleq 1 - \cos\theta - \sum_{j=1}^{m} a_j\{\cos j\theta - \cos(j-1)\theta\} \tag{24}$$

We then have

$$\sum_{n=1}^{N} E_n(X_n - X_{n-1}) = \sum_{k=1}^{N} |Z_k|^2 p_1(k2\pi/N). \tag{25}$$

We proceed in identical fashion to obtain a similar expression corresponding to the other passivity condition in Proposition 1. Note that

$$\sum_{n=1}^{N} E_n(X_n + X_{n-1}) = \mathbf{X}'\left[\mathbf{I} - \sum_{j=1}^{m} a_j \mathbf{P}^j + \mathbf{P}^{N-1} - \sum_{j=1}^{m} a_j \mathbf{P}^{j-1}\right]\mathbf{X} \tag{26}$$

Because of the previously mentioned (and crucial) property that all N

$\times N$ circulants have identical sets of eigenvectors, the diagonalizing transformation is same as the one undertaken previously in eq. (21). Hence

$$\sum_{n=1}^{N} E_n(X_n + X_{n-1}) = \sum_{k=1}^{N} |Z_k|^2 \operatorname{Re} \lambda_k \tag{27}$$

where we have denoted the eigenvalues of the matrix in eq. (26) by $\{\lambda_k\}$. Here

$$\operatorname{Re} \lambda_k = 1 + \cos\{k2\pi/N\} - \sum_{j=1}^{m} a_j[\cos\{jk2\pi/N\} + \cos\{(j-1)k2\pi/N\}] \tag{28}$$

$k = 1,2,\ldots, N$. Thus for the final form we obtain

$$\sum_{n=1}^{N} E_n(X_n + X_{n-1}) = \sum_{k=1}^{N} |Z_k|^2 p_2(k2\pi/N) \tag{29}$$

where the polynomial $p_2(\cdot)$ is defined to be

$$p_2(\theta) = 1 + \cos\theta - \sum_{j=1}^{m} a_j\{\cos j\theta + \cos(j-1)\theta\} \tag{30}$$

Now certainly Proposition 1 implies that

$$\sum_{n=1}^{N} E_n(X_n - X_{n-1}) \leqslant 0 \quad \text{and} \quad \sum_{n=1}^{N} E_n(X_n + X_{n-1}) \leqslant 0 \tag{31}$$

The above, together with eqs. (25) and (27), yields:

Proposition 2: If a periodic solution of period N with elements $(X_1, X_2, \ldots, X_N)$ exists for the recursion in eq. (1) then

$$\sum_{k=1}^{N} |Z_k|^2 p_1(k2\pi/N) \leqslant 0 \tag{32}$$

and

$$\sum_{k=1}^{N} |Z_k|^2 p_2(k2\pi/N) \leqslant 0 \tag{33}$$

where $\mathbf{Z}$, as given in eq. (21), is a transform of $\mathbf{X}$ and $p_1(\theta)$ and $p_2(\theta)$, given in eqs. (24) and (30), are two polynomials in $\cos\theta$ of degree equal to the order of the system of eq. (1).

For a fourth-order system ($m = 4$) the two polynomials are

$$p_1(\theta) = (1 + a_1) - (1 + a_1 - a_2)\cos\theta - (a_2 - a_3)\cos 2\theta - (a_3 - a_4)\cos 3\theta - a_4 \cos 4\theta \tag{34}$$

and

$$p_2(\theta) = (1 - a_1) + (1 - a_1 - a_2)\cos\theta - (a_2 + a_3)\cos 2\theta - (a_3 + a_4)\cos 3\theta - a_4 \cos 4\theta \tag{35}$$

The polynomials for second- and third-order systems are obtained from the above by setting $a_3 = a_4 = 0$ and $a_4 = 0$, respectively.

In Fig. 3a and b we have plotted $p_1(\theta)$ and $p_2(\theta)$ for a particular fourth-order system.

2.4 The main result

It is only a short step from Proposition 2 to the main result which is

Theorem 1: If for any $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$,

$$\alpha_1 p_1(\theta) + \alpha_2 p_2(\theta) > 0 \text{ for } all\ \theta\ in\ [0,\pi] \tag{36}$$

then nontrivial periodic oscillations do not exist as solutions to eq. (1).

Proof: The proof is by contradiction. Suppose a nontrivial (i.e., $\mathbf{X} \neq 0$) periodic solution of period N exists and also that the hypothesis of the theorem is valid. Then for such a solution

$$\begin{aligned}\alpha_1 \sum_{k=1}^{N} |Z_k|^2 p_1(k2\pi/N) + \alpha_2 \sum_{k=1}^{N} |Z_k|^2 p_2(k2\pi/N) \\ = \sum_{k=1}^{N} |Z_k|^2 \{\alpha_1 p_1(k2\pi/N) + \alpha_2 p_2(k2\pi/N)\} \\ > 0 \end{aligned} \tag{37}$$

from the hypothesis. However, from the passivity conditions summarized in Proposition 2,

$$\alpha_1 \sum_{k=1}^{N} |Z_k|^2 p_1(k2\pi/N) + \alpha_2 \sum_{k=1}^{N} |Z_k|^2 p_2(k2\pi/N) \leq 0 \tag{38}$$

which is a contradiction. QED.

Note that if it is desirable to know only that oscillations of a particular period N do not exist for eq. (1) then the following is a sufficient condition:

There exist

$$\alpha_1 \geq 0, \quad \alpha_2 \geq 0 \text{ such that } \alpha_1 p_1(k2\pi/N) + \alpha_2 p_2(k2\pi/N) > 0 \tag{39}$$

for $k = 1,2,\ldots,N$.

2.5 How to use the new test

Given an mth-order system, there are two simple and straightforward ways in which the above result may be used to determine whether the system does not admit overflow oscillations.

The first method requires $p_1(\theta)$, $p_2(\theta)$, and $p_1(\theta)/p_2(\theta)$ to be plotted

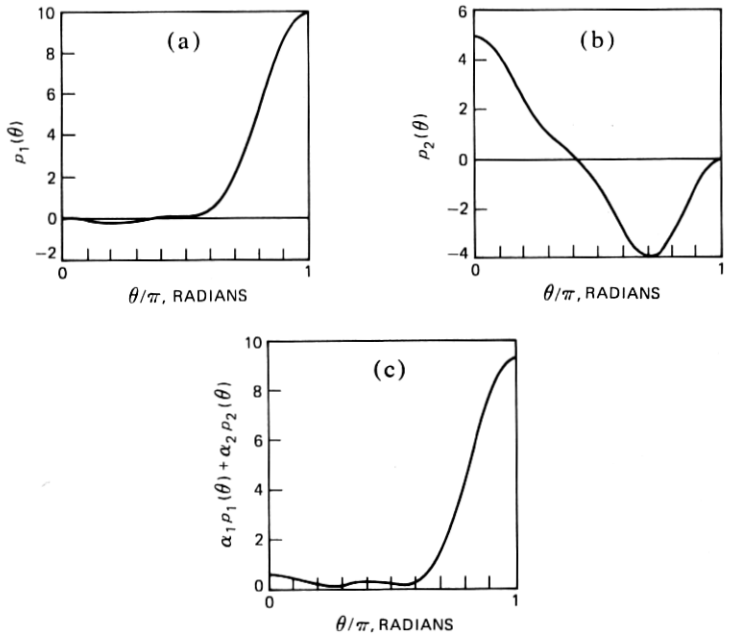

Fig. 3—Plots of polynomials $p_1(\theta)$, $p_2(\theta)$, and $\alpha_1 p_1(\theta) + \alpha_2 p_2(\theta)$ for the following fourth-order system: $a_1 = 1.1015710$, $a_2 = -1.6571120$, $a_3 = 0.7733805$, $a_4 = -0.45135546$. In (c), $\alpha_1 = .92348761$, $\alpha_2 = .16965636$

for θ in the interval $[0,\pi]$. When such plots are available the first step is to see if there is any θ for which both $p_1(\theta)$ and $p_2(\theta)$ are negative; if this is the case the hypothesis of Theorem 1 obviously cannot be satisfied and the test is automatically failed. Assuming that this is not the case we find upon reflection that the hypothesis of Theorem 1 is satisfied if and only if

$$\max_{\{\theta | p_1(\theta)>0, p_2(\theta)\leq 0\}} [p_1(\theta)/p_2(\theta)] < \min_{\{\theta | p_1(\theta)\leq 0, p_2(\theta)>0\}} [p_1(\theta)/p_2(\theta)] \tag{40}$$

In fact, if the above is true the interval defined by the left- and right-hand sides of (40) is not empty and the hypothesis of Theorem 1 is satisfied by taking $\alpha_1 = 1$ and $-\alpha_2$ to be any value in the interval.

In summary the procedure is as follows: first check to see if p_1 and p_2 are both negative at the same point. If so, then the test is failed; if not, proceed to determine the intervals where $\{p_1(\theta) > 0, p_2(\theta) \leq 0\}$ and where $\{p_1(\theta) \leq 0, p_2(\theta) > 0\}$. The test is passed (i.e., no overflow oscillations exist) if and only if the maximum of $p_1(\theta)/p_2(\theta)$ in the first interval is less than the minimum of $p_1(\theta)/p_2(\theta)$ in the latter interval.

In the second method we finely discretize the interval for θ, $[0,\pi]$, and evaluate $p_1(\theta)$ and $p_2(\theta)$ at all the discrete points $\{\theta_j\}$. Testing for the

hypothesis of Theorem 1 then amounts to testing for the consistency of the following set of linear inequalities

$$(\alpha_1\ \alpha_2)\begin{bmatrix}1 & 0 & p_1(\theta_1) & p_1(\theta_2)\ \cdots\ p_1(\pi)\\ 0 & 1 & p_2(\theta_1) & p_2(\theta_2)\ \cdots\ p_2(\pi)\end{bmatrix} \geq (1\ 1\ \ldots\ 1) \qquad (41)$$

(There is no loss of generality in assuming that the right-hand side is as specified above.) There are standard procedures[20,21] for testing for consistency of linear inequalities. In any case, Phase 1 of any commercially available linear programming package does precisely this. If a linear programming package is used then the following (dummy) functional may be used in the program: minimize $(\alpha_1 + \alpha_2)$.

The above method is easily adapted to the generalization of the criterion which is developed in Sec. VI.

III. EXAMPLES

We consider three classes of examples in some detail. In each case we tested the criterion by following the second method outlined above. We used a linear programming package (written in machine language) made available to us by A. M. Odlyzko; the interval $[0,\pi]$ was subdivided into 100 intervals. In every case the computation time was of the order of a second.

3.1 Example 1: third-order system with repeated real roots

In this class of examples we take the coefficients to depend on a real number ρ, $|\rho| < 1$, in the following manner:

$$a_1 = -3\rho, \quad a_2 = -3\rho^2, \quad a_3 = -\rho^3 \qquad (42)$$

A third-order system with the above coefficients corresponds to an underlying linear system with characteristic polynomial $(\lambda + \rho)^3$, i.e., the linear system possesses three real roots all repeated at $-\rho$. In the investigation reported in Ref. 1 we found this class of systems to be interesting for various reasons. Also, for $|\rho|$ close to 1 the behavior of the system is to some extent representative, at least with respect to oscillatory behavior, of low-pass systems and high-pass systems, depending upon whether ρ is negative or positive respectively.

In Ref. 1 we showed for system (1) that

$$|\rho| \geq 0.858 \Rightarrow \text{period-3 oscillations exist} \qquad (43)$$

Tsypkin's criterion and the circle criterion (see Sec. 5.2) give

$$|\rho| \leq 0.500 \Rightarrow \text{no overflow oscillations exist} \qquad (44)$$

An application of the new test yields

$$|\rho| \leq 0.785 \Rightarrow \text{no overflow oscillations exist}$$

Thus, in this class of examples the new test makes a substantial contribution in reducing the indeterminate region to $0.785 < |\rho| < 0.858$.

3.2 Example 2: fourth-order system with repeated real roots

The class of examples considered here is a natural extension to a higher order, $m = 4$, of the class considered in the previous example. Again all the coefficients are determined by one real parameter ρ where $|\rho| < 1$:

$$a_1 = -4\rho, \quad a_2 = -6\rho^2, \quad a_3 = -4\rho^3, \quad a_4 = -\rho^4 \tag{45}$$

Thus in this example the underlying linear system possesses four real roots, all repeated at $-\rho$.

By examining the natural set of four equations associated with a periodic solution of period 4, see eq. (12), it is easy to see that a periodic solution with elements (1,1,−1,−1) exists if and only if

$$(a_4 - a_2) \geq 1 + |a_1 - a_3| \tag{46}$$

Thus, we find on substituting for the a's that

$$|\rho| \geq 0.669 \Rightarrow \text{period-4 oscillations exist} \tag{47}$$

Tsypkin's criterion and the circle criterion give

$$|\rho| \leq 0.384 \Rightarrow \text{no overflow oscillations exist} \tag{48}$$

Application of the new criterion gives

$$|\rho| \leq 0.610 \Rightarrow \text{no overflow oscillations exist} \tag{49}$$

Thus we find that in this example too the new criterion makes an effective contribution in determining the region of stability.

3.3 Example 3: fourth-order filter for sample rate conversion

The example we consider now, a fourth-order system, was designed originally for interpolation and filtering for a terminator in a local digital switch.[4] We have reported previously[1] that in its original form the filter using saturation arithmetic sustained overflow oscillations. Here we vary one of the parameters in the design in order to estimate the modification required to guarantee the absence of oscillations. We find that the requisite variation is large. However, in the process we obtain a measure of the effectiveness of the new criterion.

The example we consider has two pairs of complex poles

$$\lambda_{1,2} = \rho_1 e^{\pm i\theta_1}, \quad \lambda_{3,4} = \rho_2 e^{\pm i\theta_2} \tag{50}$$

(The coefficients of the system are not of much interest; however, they

may be obtained from the information given below.) Also

$$\rho_1 = 0.786427817, \quad \theta_1 = 37.309784226 \text{ degrees} \tag{51}$$

and

$$\theta_2 = 39.675296075 \text{ degrees}$$

We vary ρ_2 keeping ρ_1, θ_1, and θ_2 fixed at the above values; in the original design $\rho_2 = 0.952851183$.

In (46) we have given a condition for the existence of limit cycles of period 4 with elements (1,1,−1,−1). Translating (46) to the present examples gives

$$\rho_2 \geq 0.671 \Rightarrow \text{period-4 oscillations exist} \tag{52}$$

Tsypkin's criterion and the circle criterion give

$$\rho_2 \leq 0.070 \Rightarrow \text{no oscillations exist} \tag{53}$$

An application of the new criterion gives

$$\rho_2 \leq 0.665 \Rightarrow \text{no oscillations exist} \tag{54}$$

This is a rather striking example of the effectiveness of the new criterion.

IV. AN APPLICATION

Here we examine a particular eighth-order system* which has been used in an applied research project[22] on a TDM/FDM translator.[23] The latter, a system for translating between analog frequency-division and digital time-division signals, is an extensively studied application of digital filtering. The eighth-order system has been designed to function as a low-pass filter with a sampling frequency of 8 kHz and a cutoff frequency of 2 kHz. Our object here is to demonstrate through an application of the new criterion that it is possible to design the filter as a cascade of two fourth-order sections both employing saturation arithmetic such that no overflow oscillations are sustained in either section. At least as far as overflow oscillations are concerned the margin of safety is adequate so that small changes in the coefficients due to quantization of coefficients, for example, are not going to cause overflow oscillations to appear. It should be emphasized that the result here is not a substitute for a design study and the structure suggested may well turn out to be unacceptable on grounds not related to overflow oscillations.

The system has four pairs of complex poles; the modulus (ρ_i) and

* I am grateful to V. B. Lawrence for bringing this system to my attention.

argument ($\pm\theta_i$) of each pair is as follows:

$$\rho_1 = 0.5115846\,, \quad \theta_1 = 32.870 \text{ degrees}$$

$$\rho_2 = 0.980274196, \quad \theta_2 = 80.828 \text{ degrees}$$

$$\rho_3 = 0.75259969\text{-}, \quad \theta_3 = 64.482 \text{ degrees}$$

$$\rho_4 = 0.892679\,, \quad \theta_4 = 75.297 \text{ degrees}$$

We group the first and second pairs of poles together to form one fourth-order section and the remaining pairs to form the second fourth-order section. The resulting coefficients of the two sections are, respectively,

$$a_1 = 1.1718731, \quad a_2 = -1.4912153, \quad a_3 = 0.9075846, \quad a_4 = -0.2514954 \tag{56}$$

$$a_1 = 1.1015710, \quad a_2 = -1.6571120, \quad a_3 = 0.7733805, \quad a_4 = -0.45135546 \tag{57}$$

Both sections pass the new test. For the first section it may be ascertained that with

$$\alpha_1 = 6.0819413 \text{ and } \alpha_2 = 0.07538601 \tag{58}$$

the hypothesis of Theorem 1 is satisfied. In fact, the polynomial $p_1(\theta)$ is itself positive everywhere except at $\theta = 0$, where its value is 0. However, $p_2(0) > 0$. Thus, any positive choice of α_1 and α_2 chosen suitably small will satisfy the hypothesis of Theorem 1.

For the second section (57), a choice of α_1 and α_2 for which $\alpha_1 p_1(\theta) + \alpha_2 p_2(\theta) > 0$ for all θ is

$$\alpha_1 = .92348761 \text{ and } \alpha_2 = .16965636 \tag{59}$$

Plots of $p_1(\theta)$, $p_2(\theta)$, and $\alpha_1 p_1(\theta) + \alpha_2 p_2(\theta)$ for the second section are displayed in Fig. 3.

It is noteworthy that both sections fail Tsypkin's criterion and the circle criterion.

V. SOME IMPLICATIONS OF THE MAIN RESULT (THEOREM 1)

5.1 Overflow oscillations do not exist in second-order systems

It is well known[12,13,14] that when the order of the system in eq. (1) is two, then overflow oscillations are not sustained. The proofs of this result are rather special to second-order systems and to the saturation arithmetic. On the other hand, there are the frequency-domain criteria[5–11] for stability which are systems-theoretical results applicable to large classes of nonlinearities and systems of arbitrary order. However, we may

infer from the results in Ref. 8 that these criteria do not give the result that all second-order systems are free from overflow oscillations.

We show that the criterion in Theorem 1 does give the well-known result on second-order systems. Our result is given in Proposition 3. [It is assumed that $|a_2| < 1$ and $1 - |a_1| - a_2 > 0$; these relations are equivalent to eq. (3), i.e., the underlying linear system is stable.]

Proposition 3: Let $m = 2$ in eq. (1). Also let

$$\alpha_1 = (1 + a_1 - a_2) > 0 \text{ and } \alpha_2 = (1 - a_1 - a_2) > 0 \tag{60}$$

Then,

$$\alpha_1 p_1(\theta) + \alpha_2 p_2(\theta) > 0 \text{ for } all\ \theta \tag{61}$$

The proof of this result is in Appendix B. The above in conjunction with Theorem 1 shows that oscillations are not sustained in second-order systems.

5.2 Tsypkin's criterion and discrete circle criterion are subsumed by new criterion

The object here is to show that the new criterion subsumes both Tsypkin's criterion[5] and the discrete circle criterion[11] when the latter criteria are used to determine the nonexistence of oscillations in eq. (1). The two closely related frequency-domain criteria are identical when applied to the system in eq. (1).

Tsypkin's criterion[5] is as follows in applications to systems like eq. (1) where the nonlinearity F satisfies

$$K_{\min} \leqslant F(\sigma)/\sigma \leqslant K_{\max} \quad \text{for } all\ \sigma \tag{62}$$

If

$$(i)\ \sum a_j z^{-j}/[1 - K_{\min} \sum a_j z^{-j}] \text{ is finite for } all\ |z| \geqslant 1 \tag{63}$$

$$(ii)\ \frac{1}{K_{\max} - K_{\min}} - \text{Re}\,[\sum a_j e^{-ij\theta}/(1 - K_{\min} \sum a_j e^{-ij\theta})] > 0 \text{ for } all\ \theta \text{ in } [0,2\pi] \tag{64}$$

then $\lim_{n\to\infty} x_n = 0$; in particular, oscillations do not exist.

In the case of eq. (1) where F is the saturation nonlinearity,

$$K_{\min} = 0 \quad \text{and} \quad K_{\max} = 1 \tag{65}$$

so that the effective restriction is (64) which reduces to

$$1 - \sum_{j=1}^{m} a_j \cos j\theta > 0 \quad \text{for } all\ \theta \tag{66}$$

When (65) holds the discrete time version of the circle criterion[11] is identical to the above condition.

In Theorem 1 let $\alpha_1 = \alpha_2 = 1/2$. Then from the defining relations for $p_1(\theta)$ and $p_2(\theta)$ in (24) and (30) respectively, we find that

$$\alpha_1 p_1(\theta) + \alpha_2 p_2(\theta) = 1 - \sum_{j=1}^{m} a_j \cos j\theta \tag{67}$$

Thus, as previously asserted, if either Tsypkin's criterion or the circle criterion is satisfied, i.e., (66) is valid, then the hypothesis of Theorem 1 is also satisfied.

VI. A GENERALIZATION OF THE MAIN RESULT

The reader will recall that the main result, Theorem 1, is a direct consequence of the rather special passivity properties, stated in Sec. 2.1, which are implied by the special features of the saturation nonlinearity F. Another key ingredient is that the passivity conditions imply inequalities on quadratic forms involving circulants. We show here that many conditions akin to the ones in Proposition 1 are valid by virtue of the properties of the saturation nonlinearity. All or some of these may be used to augment the passivity conditions used so far so as to obtain improved criteria for the nonexistence of oscillations.

The following generalized passivity conditions exist* for any $l \geq 1$:

$$e_n(x_n - x_{n-l}) \leqslant 0 \quad n = l, l+1, \ldots \tag{68}$$

$$e_n(x_n + x_{n-l}) \leqslant 0 \quad n = l, l+1, \ldots \tag{69}$$

where $\{x_n\}$ is any solution of eq. (1) and $\{e_n\}$ is obtained from the solution through eq. (6). The proof is similar to that of Proposition 1. Thus in Proposition 1 we have used only a very small subset ($l = 1$) of all the above conditions.

The interesting fact is that each of the expressions in the above conditions summed over N, where N is the period of any periodic solution of (1), is equivalent to a quadratic form involving a circulant. Thus if $\mathbf{X} = (X_1, X_2, \ldots, X_N)'$ are the elements of the periodic solution and $(E_1, E_2, \ldots, E_N)$ are the corresponding forcing terms, see eqs. (12) and (13), then

$$\sum_{n=1}^{N} E_n(X_n - X_{n-l}) = \mathbf{X}' \left[\mathbf{I} - \sum_{j=1}^{m} a_j \mathbf{P}^j - \mathbf{P}^{-l} - \sum_{j=1}^{m} a_j \mathbf{P}^{-l+j} \right] \mathbf{X} \tag{70}$$

* The generalized passivity conditions are also valid for negative values of l although we do not make any explicit use of this fact.

$$\sum_{n=1}^{N} E_n(X_n + X_{n-l}) = \mathbf{X}' \left[\mathbf{I} - \sum_{j=1}^{m} a_j \mathbf{P}^j + \mathbf{P}^{-l} + \sum_{j=1}^{m} a_j \mathbf{P}^{-l+j} \right] \mathbf{X} \tag{71}$$

for $l = 1,2,\ldots$. Hence by transforming $\mathbf{X}$ to $\mathbf{Z}$ where $\mathbf{Z} = \mathbf{U}^*\mathbf{X}$, $\mathbf{U}$ being the unitary matrix of eigenvectors of $N \times N$ circulants, we obtain for $l = 1,2,\ldots$

$$\sum_{n=1}^{N} E_n(X_n - X_{n-l}) = \sum_{k=1}^{N} |Z_k|^2 p_l(k2\pi/N) \tag{72}$$

and

$$\sum_{n=1}^{N} E_n(X_n + X_{n-l}) = \sum_{k=1}^{N} |Z_k|^2 p'_l(k2\pi/N) \tag{73}$$

where

$$p_l(\theta) \triangleq 1 - \cos l\theta - \sum_{j=1}^{m} a_j\{\cos j\theta + \cos (j-l)\theta\} \tag{74}$$

and

$$p'_l(\theta) \triangleq 1 + \cos l\theta - \sum_{j=1}^{m} a_j\{\cos j\theta - \cos (j-l)\theta\} \tag{75}$$

Thus $p_1(\theta)$ and $p_2(\theta)$ defined in Sec. 2.3 correspond to $p_1(\theta)$ and $p'_1(\theta)$ respectively in the present notation.

Certainly the generalized passivity condition in (68) and (69) imply that the expressions in (72) and (73) are nonpositive. We thus arrive at the following generalization of Theorem 1:

Theorem 2: If any convex linear combination of $p_1(\theta), p'_1(\theta), p_2(\theta), p'_2(\theta), \ldots$ is positive for all θ in $[0,\pi]$, then the system in eq. (1) does not have any nontrivial periodic solutions.

In experiments involving fourth-order systems of practical interest we have not found the use of the above generalized criterion to make any material difference in delineating the stable systems. In these investigations we used a linear programming package (Sec. 2.5) to apply the test in Theorem 2 with up to six polynomials (the leading six polynomials of Theorem 2) being used. However, it is quite possible for substantial improvements to exist in other cases.

ACKNOWLEDGMENTS

We gratefully acknowledge our debt to A. M. Odlyzko for making available and in assisting us in the use of a linear programming package. We are grateful to V. B. Lawrence for bringing the filter in Section IV to our attention.

APPENDIX A

Circulant matrices

For completeness we collect here some of the well-known properties of circulants which are used in the paper. The interested reader may refer to Muir[16] and Grenander and Szego[17] for further details and applications; Ref. 15 concisely lists some of the main properties.

We let $\mathbf{P}$ denote the primitive $N \times N$ circulant:

$$\mathbf{P} = \begin{bmatrix} 0 & — & — & — & — & 0 & 1 \\ 1 & 0 & — & — & — & — & 0 \\ 0 & 1 & 0 & — & — & — & 0 \\ — & — & — & — & — & — & — \\ 0 & — & — & — & — & 1 & 0 \end{bmatrix} \tag{76}$$

Note that

$$\mathbf{P}^N = I \tag{77}$$

and that

$$\mathbf{P}' = \mathbf{P}^{N-1} = \mathbf{P}^{-1} \tag{78}$$

A polynomial of arbitrary degree in $\mathbf{P}$ is a circulant. An $N \times N$ circulant $\mathbf{C}$,

$$\mathbf{C} = \sum_{j=0}^{N-1} c_j P^j \tag{79}$$

has as its eigenvalues

$$\sum_{j=0}^{N-1} c_j e^{-ijk2\pi/N} \quad k = 1,2,\ldots,N \tag{80}$$

All $N \times N$ circulants have as eigenvectors $\mathbf{u}_k$, $k = 1,\ldots,N$, given in eq. (16). The matrix $\mathbf{U}$ with the eigenvectors as columns is unitary, i.e.,

$$\mathbf{U}^*\mathbf{U} = I \tag{81}$$

APPENDIX B

Proof of proposition 3

We prove here the assertion in Proposition 3, namely, for second-order systems

$$q(\theta) \triangleq (1 + a_1 - a_2)p_1(\theta) + (1 - a_1 - a_2)p_2(\theta) > 0 \quad \text{for } all\ \theta \tag{82}$$

For second-order systems

$$p_1(\theta) = (1 + a_1) - (1 + a_1 - a_2)\cos\theta - a_2 \cos 2\theta \tag{83}$$

and,

$$p_2(\theta) = (1 - a_1) + (1 - a_1 - a_2)\cos\theta - a_2\cos 2\theta \tag{84}$$

We find upon substitution that

$$q(\theta) = -4a_2(1 - a_2)\cos^2\theta - 4a_1(1 - a_2)\cos\theta + 2(1 + a_1^2 - a_2^2) \tag{85}$$

First observe that at the corner points q is positive:

$$q(0) = 2(1 - a_1 - a_2)^2 > 0 \text{ and } q(\pi) = 2(1 + a_1 - a_2)^2 > 0 \tag{86}$$

Through differentiation we find that minima of $q(\theta)$ occur in the interior of the region $[0,\pi]$ if and only if

$$|a_1| \leqslant -2a_2 \tag{87}$$

and that at a minimum $\hat{\theta}$,

$$\cos\hat{\theta} = -a_1/2a_2 \tag{88}$$

Evaluating q at such a point we obtain

$$q(\hat{\theta}) = \frac{-(1 + a_2)}{a_2}[-2a_2(1 + a_2) + (4a_2^2 - a_1^2)] \tag{89}$$

$$> 0$$

REFERENCES

1. Debasis Mitra, "Large Amplitude, Self-Sustained Oscillations in Difference Equations Which Describe Digital Filter Sections Using Saturation Arithmetic," IEEE Trans. Acoustics, Speech, Signal Proc., April 1977.
2. A. Croisier, D. J. Esteban, M. E. Levilion and V. Rizo, "Digital Filter for PCM Encoded Signals," U. S. Patent 3777130, December 3, 1973.
3. A. Peled and B. Liu, "A New Hardware Realization of Digital Filters," IEEE Trans. Acoust., Speech, Signal Proc., *ASSP-22,* December 1974, pp. 456–462.
4. R. B. Kieburtz, "Interpolation and Filtering for the T-Line Terminator," Bell Laboratories internal memorandum, March 1975. Also, "Digital Interpolation Interface Between Two Systems at Slightly Different Sampling Rates," presented at IEEE Workshop on Signal Processing, Arden House, 1976.
5. Ya. Z. Tsypkin, "Fundamentals of the Theory of Non-Linear Pulse Control Systems," Proc. Second Intl. Cong., Intl. Fed. of Automatic Control, Basel, 1963, pp. 172–180.Also, "A Criterion for Absolute Stability of Automatic Pulse-Systems with Monotonic Characteristics of the Non-Linear Element," Sov. Phys. Dokl., *9,* October 1964, pp. 263–266.
6. E. D. Garber, "Frequency Criteria for the Absence of Periodic Responses," Automat. Remote Contr., *28,* No. 11, November 1967.
7. A. I. Barkin, "Sufficient Conditions for the Absence of Auto-Oscillations In Pulse Systems," Automat. Remote Contr., *31,* June 1970, pp. 942–946.
8. T. Claasen, W. F. G. Mecklenbraüker, and J. B. H. Peek, "Frequency Domain Criteria for the Absence of Zero-Input Limit Cycles in Nonlinear Discrete-Time Systems, With Applications to Digital Filters," IEEE Trans. Circuits Syst., *CAS-22,* No. 3 (March 1975) pp. 232–239.
9. I. W. Sandberg, "On the Boundedness of Solutions of Nonlinear Integral Equations," B.S.T.J., *44,* No. 3 (March 1965), pp. 439–453.
10. G. P. Szego, "On the Absolute Stability of Sampled-Data Control Systems," Proc. Nat. Acad. Sci., *50,* 1963, pp. 558–560.
11. J. L. Willems, *Stability Theory of Dynamical Systems,* New York: John Wiley, 1970; Ch. 6.

12. P. M. Ebert, J. E. Mazo, and M. G. Taylor, "Overflow Oscillations in Digital Filters," B.S.T.J., *48*, No. 9 (November 1969), pp. 2999–3020.
13. I. W. Sandberg, "A Theorem Concerning Limit Cycles in Digital Filters," Proc. 7th Ann. Allerton Conf. Circuits and Systems Theory, pp. 63–68, 1969.
14. A. N. Willson, Jr., "Limit Cycles Due to Adder Overflow in Digital Filters," IEEE Trans. Circuit Theory, *CT-19*, 1972, pp. 342–346.
15. M. Marcus, "Basic Theorems in Matrix Theory," Nat. Bur. Stds., Applied Math. Series 57, January 1960, p. 9.
16. T. Muir, *A Treatise On The Theory of Determinants*, Dover Publications, New York, 1960; Ch. 12.
17. U. Grenander and G. Szego, *Toeplitz Forms And Their Applications*, Berkeley: Univ. of Calif. Press, 1958; Ch. 8.
18. Debasis Mitra, "A Bound on Limits in Digital Filters which Exploits a Particular Structural Property of the Quantization." A summary appears in Proc. 1977 IEEE Intl. Conf. on Acoustics, Speech, and Signal Proc., May 1977. The full paper is due for publication in IEEE Trans. Circuits Syst., Nov. 1977.
19. A. N. Willson, Jr., "Computation of the Periods of Forced Overflow Oscillations in Digital Filters," IEEE Trans. Acoustics, Speech, Signal Proc., *ASSP-24*, No. 1 (February 1976).
20. D. Gale, "How to Solve Linear Inequalities," Amer. Math. Monthly, *76*, 1969, pp. 589–599.
21. G. Debreu, "Non-negative Solutions of Linear Inequalities," Internat. Economic Review, Vol. 5, 1964.
22. R. B. Kieburtz, V. B. Lawrence, and K. V. Mina, "Control of Limit Cycles in Recursive Digital Filters by Randomized Quantization," Bell Laboratories memorandum; also talk presented at Intl. Symp. Circuits and Systems, Munich, Germany, April 1976.
23. S. L. Freeny, R. B. Kieburtz, K. V. Mina, and S. K. Tewksbury, "Design of Digital Filters for an all Digital Frequency Division Multiplex-Time Division Multiplex Translator," IEEE Trans. Circuit Theory, *CT-18* No. 6 (November 1971).
24. S. Lefschetz, *Stability of Nonlinear Control Systems*, Academic Press, New York, 1965; Ch. 1 and 2.